International Statistical Review

International Statistical Review (2015), 0, 0, 1–22 doi:10.1111/insr.12110

Implications of the Data Revolution for Statistics Education

Jim Ridgway

School of Education, University of Durham, Leazes Road, Durham DH1 1TA, UK
E-mail: jim.ridgway@durham.ac.uk

Summary

There has never been a more exciting time to be involved in statistics. Emerging data sources provide new sorts of evidence, provoke new sorts of questions, make possible new sorts of answers and shape the ways that evidence is used to influence policy, public opinion and business practices. Significant developments include open data, big data, data visualisation and the rise of data-driven journalism. These developments are changing the nature of the evidence that is available, the ways in which it is presented and used and the skills needed for its interpretation. Educators should place less emphasis on small samples and linear models and more emphasis on large samples, multivariate description and data visualisation. Techniques used to analyse big data need to be taught. The increasing diversity of data usage requires deeper conceptual analysis in the curriculum; this should include explorations of the functions of modelling, and the politics of data and ethics. The data revolution can invigorate the existing curriculum by exemplifying the perils of biassed sampling, corruption of measures and modelling failures. Students need to learn to think statistically and to develop an aesthetic for data handling and modelling based on solving practical problems.

Key words: statistics education; modelling; open data; big data; visualisation; data-driven journalism; curriculum; statistical literacy; change.

1 Introduction

Developments enabled by technology are shaping the ways that evidence is used to influence public opinion and policy. The open data movement aims to make high quality data collected by governments and non-governmental organisations accessible to citizens. Big data—ill-structured and opportunistic data derived from sources such as the location of mobile phones or analysis of traffic on e-mail—are being used more and more, often for purposes unforeseen by the originators of the data. Increasingly, journalists are making good use of rich data and are creating excellent, data-laden websites. An exciting range of visualisations is being developed to allow users to explore large, rich data sets. It is reasonable to assert that the whole 'knowledge landscape' is being transformed. New sorts of data are available (for example, semantic analyses of *twitter* streams), and familiar sorts of data can be accessed and explored in new ways (for example, the *constituency explorer* offers dynamic visualisations of census data via mobile devices). The data revolution provokes further reflections on the essential nature of both 'statistics' and 'statistics education' (Ridgway, Nicholson & McCusker, 2011, 2013).

There is a need to rebalance the statistics curriculum. Big data and open data make the need for statistical thinking even more important and provide contexts for discussing core statistical ideas such as sampling bias and causality. Much of the statistics curriculum in many countries

can be traced to the 1930s (Batanero *et al.*, 2011); there is a need to extend the range of models that is taught, and to discuss the nature of modelling and ways to validate models. Statistics is central to discourses about evidence and policy; discourses about evidence and policy should be central in the statistics curriculum.

2 The Data Revolution

2.1 Open Data

As early as 1792, Condorcet (1994) asserted the importance of informing citizens about governance and presenting evidence about the state of society, in order to increase awareness of injustices and structural social inequalities. He believed in *savoir liberateur*—knowledge that would enable people to free themselves from social oppression. More recent initiatives such as data.gov in the USA and data.gov.uk in the UK explicitly state political objectives, notably to promote the democratic process by giving citizens access to data that can stimulate debate and inform policy making. Economic advantage is a second driver. For example, in the UK, the Open Data Institute claims that it 'will catalyse the evolution [of the] open data culture to create economic, environmental, and social value. It will unlock supply, generate demand, create and disseminate knowledge to address local and global issues'. The open data movement has had considerable success in recent years in persuading major data providers (such as national statistics offices, Eurostat and governments) to make data available to anyone who wants it. The promises of open data are obvious—but there are barriers to be overcome. 'Open' does not mean 'readily accessible'—users often need to be familiar with extensible markup language (XML), JavaScript object notation (JSON) and using application programming interfaces (API) to access data, as well as with excel or other proprietary software such as `SPSS` or `SAS`. Synthesising data from different sources requires access to, and understanding of, metadata. Interpreting multivariate data is non-trivial. The task of drawing conclusions from samples to populations is different to the task of drawing conclusions about sub-samples from populations. Statistics educators are faced with the challenge of educating an entire population about seemingly difficult ideas. In the classroom, there is a need to create curricula that devote more attention to the interpretation of large scale data sets.

2.2 Big Data

'Big data' refer to a variety of sources and data sets that have a number of characteristics in common, but where different examples do not exhibit identical characteristics—rather like the word 'game' that can describe both solitaire and soccer. 'Big data' usually refer to data that arise naturally from some system and are a by-product of that system ('exhaust data'); data are often collected in real time; the volume is torrential and cannot be stored, managed and analysed via conventional methods, and the signal variance is very high (and often associated with high levels of noise). Examples include data from sensors (e.g. from wearable appliances and meteorological stations), transactional data (e.g. from mobile phones and supermarket transactions) and data scraped from web pages (e.g. when using price changes to assess inflation). 'Undoubtedly the greatest challenge and opportunity that confronts today's statisticians is the rise of Big Data' (Madigan, 2014, p4). Big data differ from open data in a number of critical respects. Open data have been created for some purpose; measures are clearly defined; data are usually multivariate; the population being sampled is known, and the whole process of data generation and presentation has been subjected to extensive scrutiny. It is usually possible to gauge the robustness of the evidence and the extent to which data might be misleading for

innocent or nefarious reasons. Big data often have none of these characteristics; further, they are often not 'open' but are owned by companies who seek economic advantage from their use. Big data are particularly interesting to social scientists because new phenomena are measured, and these measures are usually measures of behaviour, not internal states mediated by verbal reports (e.g. attitudes, opinions, beliefs and memories) that characterise much of the large scale survey data available.

As in the case of open data, there are challenges and opportunities for statistics educators. Big data pose new problems for acquisition, storage and subsequent access. Data acquisition techniques are interesting in themselves such as recording data only in response to trigger events (as in particle physics), and data often require a great deal of pre-processing (e.g. data from sensors such as satellites). There are threats to the whole notion of 'statistician'—computer science skills are essential for accessing, storing and processing big data; traditional methods developed by statisticians for data storage, access and analysis cannot be applied. One can ask what added value 'statisticians' actually bring and how collaborations can be forged with computer scientists.

For statistics educators, the skills base of teachers and students needs to be extended to include an understanding of the analytic techniques suited to accessing, storing (perhaps) and analysing high-volume unstructured data. Big data can provide interesting contexts to explore fundamental ideas such as sampling, data quality, the principles of measurement and plausible inference in the face of uncertainty.

2.3 Data Visualisation and Data-Driven Journalism

Visual representation of data has been going on for at least 1 000 years (http://datavis.ca/milestones/). Graphics have a long history of being used to make social data accessible to a wide audience, exemplified by the Neuraths' work in the period 1930–1945 designing graphical displays to demonstrate social inequalities (Neurath, 2010). Major data providers are providing visualisations to make their data more accessible (e.g. Organisation for Economic Cooperation and Development data are accessible via Gapminder and eXplorer). Key political targets such as the United Nations (UN) Millennium Development Goals (MDG) are presented in the form of an interactive dashboard, to encourage public engagement. As with 'big data', 'data visualisation' describes quite different conglomerations of features. Users can have varying degrees of control; displays may be updated automatically or not. For example, http://earth.nullschool.net/ shows global wind movement derived from sensors; the user can only control the viewpoint taken and interrogate points on the globe. Http://bit.ly/1BIx5zE (Figure 4) shows counts of the first incidences of sexually transmitted diseases (STD) as a function of sex, age, disease and time; the user can decide what to plot and what to fix and can slide time. There is no automatic updating of data from the health database. All techniques for data presentation and analysis (including standard statistical packages) make some explorations easy and some rather difficult. All data visualisations limit users' choices (e.g. of data sources, combinations of variables, comparisons that can be made and inferences that can be drawn within the display). They provide a rich resource for statistics educators to discuss the strengths and weaknesses of different data representations and exploratory tools.

As well as a plethora of new ways to present data, an important cultural trend is the emergence of data-driven journalism (e.g. Bradshaw, 2010; Brooke, 2010; Gray, Chambers, & Bounegru, 2012). Early, influential examples of data-driven journalism in the UK can be found in the coverage of the *Wikileaks* cables. Journalists had to work with large data sets, adopting crowdsourcing techniques to extract and publish high profile stories. In content terms, data journalism is typified by visually rich, often interactive, articles. Infographics feature heavily; text is often abbreviated to support the visuals.

Gal (2002) argues that statistical literacy refers to 'the need for people ... to develop the ability to comprehend, interpret, and critically evaluate messages with statistical elements or arguments conveyed by the media and other sources'. It follows that statistics education needs to keep up with the increasing sophistication of journalists. Learners need to be equipped with skills to interpret new data visualisations and to critique increasingly sophisticated, data-rich arguments.

3 Statistical Thinking and Doing Statistics

Pullinger (2014) offers biographies of the first committee of the Royal Statistical Society that included Henry Lansdowne (50 years a politician, including a role as Chancellor of the Exchequer); Charles Babbage (mathematician, engineer, astronomer and inventor of the computer); John Elliot Drinkwater (administrator and champion of girls' education in India); Henry Hallam (historian and political activist) and Richard Jones (applied economist). This gathering reflects the diverse but highly practical background of the founding members. The development of statistics in the late 19th and early 20th century was characterised by the invention of branches of mathematics in response to practical challenges in biology and agriculture. The early history provides evidence of engagement with practical sciences that includes the design and analysis of studies to investigate and improve crop yields (with its linguistic legacy via terms such as 'spit plot designs') and Nightingale's dramatic visualisation of deaths due to combat and to disease during the Crimean war.

Tukey (1962) argued that statistics education should focus more on problems and less on teaching how to use tools; statisticians should seek out problems that 'offer unusual challenges' (p3) and should 'take up the rock road of real problems' (p64). The Royal Statistical Society has the strap line 'data, evidence and decisions', and their website (Royal Statistical Society, 2014) asserts 'Statistics is fundamentally about information, numerical data, and about applying quantitative skills to real problems'. Hahn and Doganaksoy (2011, cited on the World of Statistics website http://www.worldofstatistics.org/) describe statistics as 'the quintessential interdisciplinary science; and the art of telling a story with [numerical] data'. Hand (2009) asserts that 'Statistics is about solving real problems. An undue emphasis on its mathematical foundations is detrimental to the discipline'. It follows that the discipline of statistics should evolve in response to the changes in the sorts of problems to be addressed and to the sorts of data that are available. Seismic changes are now required, rather than fine adjustments. In Pullinger's words (Pullinger, 2014, p821) 'current directions call us to invest in new thinking about our idea of statistics and how it makes an impact'. Statistics education needs to reflect the dramatic changes in the data sources available, the ways they are accessed, stored and analysed and the uses to which they are put; modelling phenomena should be at the core of the curriculum.

4 Models and Modelling

'Most real life statistical problems have one or more non-standard features. There are no routine statistical questions; only questionable statistical routines.' (Cox, quoted in Chatfield, 1991, p240). Scientists (including social scientists) and engineers can view statistics as a service subject, where all the mathematics has been created and the role of the statistician is to find 'the' model that fits the data. Application of standard models needs to be performed with care. 'All models are wrong, but some are useful' (Box & Draper, 1987, p424). Consider perhaps the simplest model—the use of a summary statistic. George W. Bush introduced tax cuts, claiming that 92 million Americans would receive an average tax cut of $1083 in 2003

(US Congress, 2004, p22). What was not pointed out was that the median tax cut would be about $100. Similarly, modelling events that have leptokurtic distributions such as droughts, rainfall or sea waves with a normal distribution can have unfortunate consequences (e.g. Zetie & James, 2002). Over-dependence on standard models (e.g. making assumptions about linearity and continuous functions) ignores the history of statistics; new situations often require the invention of new mathematical structures and methods. The mathematical sciences in general, and statistics in particular, have consistently created transformative theory and methods when faced with interesting problems. Engagement with contemporary problems has been a hallmark of statistics; this should be reflected in the curriculum.

Kuhn's notion (Kuhn, 1962) of a paradigm shift is relevant, here. A 'paradigm' is a world view that maps out an area of interest, some theoretical assumptions, agreed facts (veridical or not), discovery methods, methods for analysis and allowable conclusions. Kuhn's ideas (Kuhn, 1962) are exemplified by the different approaches taken to evidence in economics, psychology and epidemiology (Grolemund and Wickham (2014) go so far as to argue that statistics has been balkanised). In statistics education, the dominant paradigm can be conceptualised (with a few exceptions) as teaching context-independent methods of data analysis (or using data sets collected for a specific purpose, which is perfectly aligned with the question asked); use of data sets that are small enough to fit into standard statistical packages; applying linear models where the elements are explicit; assuming hierarchical definitions of subject difficulty (with univariate analyses judged to be inherently easier than multivariate analyses) and adopting numerical analysis as the primary tool for data exploration and analysis. Every aspect of this paradigm is challenged by the data revolution. A new set of problems is emerging that requires custom-designed methods; data may be available that are plausibly relevant to a question of interest, but are not an exact fit; data sets are often very large; a critical activity may well be to identify and define key variables; linear models may not be useful (in much of school science, for example); the difficulty of different concepts for learners is ill-understood (e.g. Ridgway, McCusker & Nicholson (2003) use Rasch scaling to show that computer-based problems involving three variables can be easier (in psychometric terms) than one and two variable problems presented on paper); visual methods can be very powerful analytic tools.

Perhaps the most radical (and strongly contested) claim for the data revolution was offered by Anderson (2008) in a Wired Magazine editorial entitled The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. 'Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves'. Anderson (2008) gives examples of shotgun-gene sequencing that has led to the discovery of new species and language translation via pattern matching rather than linguistic analysis. The notion that systems (e.g. in engineering design) that cannot be described algebraically can be modelled by numerical methods is quite familiar. Widening the challenge to science and social science is an important provocation to reflect on the nature of modelling and the ways it should be addressed in the statistics curriculum. The most conservative conclusion to be drawn is that the nature and purpose of modelling *should* be addressed; students should learn to critique different models of phenomena.

It is also clear that the curriculum should encompass some of the techniques used in the analysis of big data. Statistics educators should also grasp the opportunities that open data offer as a royal road into social science. Open data are directly relevant to grounded theory. 'Grounded theory . . . is a systematic methodology in the social sciences involving the discovery of theory through the analysis of data . . . the first step is data collection . . . This contradicts

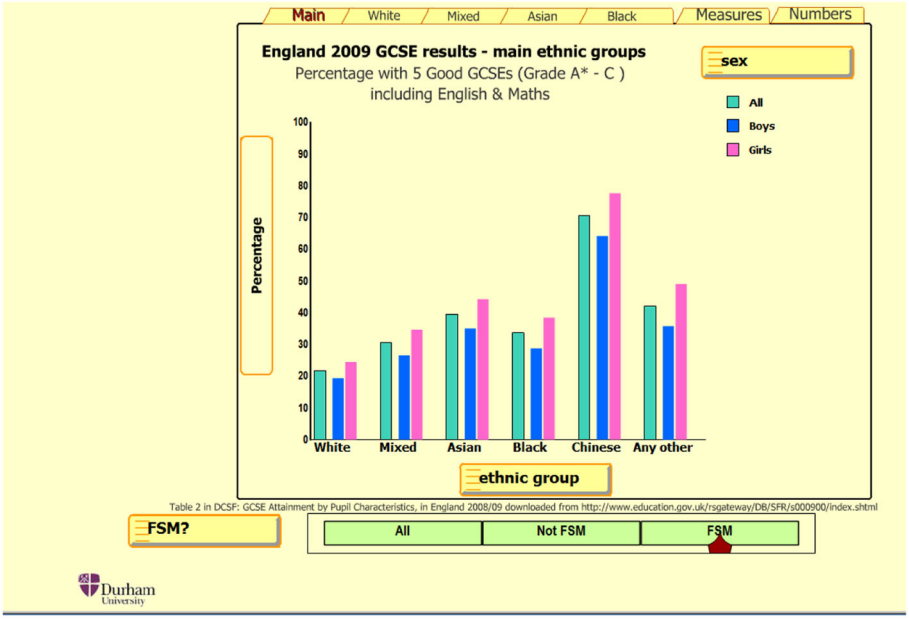

Figure 1. *Educational attainment of students in receipt of free school meals (5 or more General Certificate of Secondary Education passes).*

the traditional model of research, where the researcher chooses a theoretical framework, and only then applies this model to the phenomenon to be studied' (Wikipedia, 2014). It might seem extraordinary to statisticians that there is any alternative to grounded theory, and even more extraordinary that grounded theory 'contradicts the traditional model of research'. Some statisticians might not regret the end of 'the traditional model' portrayed here. The data in Figure 1 can illustrate the contrasting styles of theorising and the usefulness of open data. Figure 1 shows data on educational attainment at the end of compulsory schooling on national tests by students who are 16 years old and eligible for free school meals (FSM) in England and Wales in 2010, broken down by sex and ethnic group. Variable names ('sex', 'ethnic group' etc.) can be dragged to different locations; sliders reveal different subsets of the data ('All', 'Not FSM' and'FSM') to facilitate data exploration. Different data sets are provided under different tabs ('White', 'Mixed' etc). Notice that girls in every group outperform boys; white children eligible for FSM perform *worse* than eligible black children, and Chinese students outperform others by a considerable margin.

There is a number of theories about the factors that underpin educational attainment, none of which are compatible with the data in Figure 1. For example, beliefs in 'intelligence' have been used to justify selective education. To fit the data, one needs to explain the superiority in attainment of all Chinese students, of girls and of (Chinese excluded) students not eligible for FSM. In contrast, starting with the raw data in Figure 1 provokes questions about both theory and its practical implications. Why do girls in every group outperform boys? Why do white children eligible for FSM perform worse that black children (given that the reverse is true for pupils not eligible for FSM)? and Why do Chinese students outperform others by a considerable margin (irrespective of being eligible for FSM or not)?

The data revolution provokes discussions about the nature and purpose of theorising and modelling. Scientific disciplines differ in the emphasis they place on understanding, prediction and controlling future events. Newton's laws of motion allow predictions to be made and

facilitate control over some future events (such as landing on the moon); they require no understanding of the nature of 'gravity' or 'mass' for them to be of practical use. In some disciplines, data are gathered primarily for better understanding. Astronomy provides an example; the primary ambition is not to improve the universe, nor to use it for some commercial purpose (but see http://www.asterank.com/). In some disciplines such as psychology and engineering, there are ambitions to understand, predict and control future events and behaviours. Open data and big data differ in the extent to which they provide answers to different questions. For example,

What is happening now? Open data and data visualisation can map out the phenomena to be explored, as in the example of educational attainment data at the end of compulsory schooling in England and Wales. Big data can be used to show live streaming of traffic flows, weather systems or tweets. These data can be viewed as atheoretical (although the choice of measures in open data has theoretical underpinnings and attention to certain big data measures reflects a belief that they are important) or pre-theoretical. Describing phenomena is a starting point in any scientific investigation.

What will happen next? Tomorrow? In the future? Models are used for prediction. Uses of big data for prediction often assume that past patterns will be repeated in the future. In the simplest case, if two products (e.g. books) were often bought together by people in the past, they are likely to be bought together in the future. In the case of neural net models of credit worthiness, there is an assumption that credit-worthy people have identifiable sets of characteristics that can be used to make decisions about loans; in the case of Google Flu Trends (GFT) (discussed later), there was an assumption that things that predict flu epidemics at one time will predict flu epidemics in the future. The data revolution provokes questions about ways to predict, ways to evaluate rival accounts and predictive power. Silver (2012) provides interesting insights into a variety of predictive models. One example was an exploration of the predictive power of different weather forecasting models. The simplest was *persistence*—the assumption that the weather tomorrow will be the same as it is today; next was *climatology*—the assumption that tomorrow's weather will reflect the long-term average of weather in a particular place on a particular day; the third was weather forecasts using sophisticated computer-based models of the atmosphere. Atmosphere models predicting temperature made better predictions for forecasts 1 to 8 days ahead then were *worse* than climatology models (p132)—past events predicting future ones.

From the viewpoint of statistics education, the data revolution provokes discussions about the ways that data are being used to make decisions that affect everyday life and about a wide range of predictive models. The idea of prediction is conceptually simpler than hypothesis testing and should be introduced early into the curriculum. Students should be introduced to ideas of prediction, different strategies (and models) for prediction and ways to evaluate them.

What would happen if we changed things this way or that way (e.g. decisions about social policy)? This is the area of both theory and ideology, where prior beliefs and existing models are used to make predictions around 'what if' conjectures underpinned by 'what ought' beliefs. Complex systems such as social systems or the environment are characterised by feedback and feed forward loops; not only are the effects of changes difficult to predict but so too is the reaction to specific models. Meadows *et al.* (1972) *Limits to Growth* identified and illustrated the problems of assuming exponential growth when modelling complex systems. The controversies it aroused highlight the problems of ideologies that lock systems in stasis and prevent appropriate political action (e.g. Bardi, 2008). Starting from educational theories about the underpinnings of educational attainment, or the data in Figure 1, would lead to quite different educational policies. Single variable accounts ('intelligence') discourage action to boost educational attainment; the data suggest that actions should be explored that relate to home and school cultures, as well as poverty. Reports of the death of theory are much exaggerated.

5 Problems with Statistics Curricula

Statistical investigations have a number of components (shared with applied sciences) that include problem definition and redefinition; problem representation and re-representation; defining and refining measures; data collection and cleaning; modelling data and modelling phenomena; estimating the probabilities of different events, given different interventions (or inaction) and estimating likely costs and benefits associated with different actions. A number of authors (Royal Statistical Society, 2014; Wild & Pfannkuch, 1999) refer to a statistics cycle; others (e.g. Huber, 2011) resist the idea of a cycle (because the idea of a cycle implicitly restricts the interconnectivity of all the components). From the viewpoint of curriculum design, it is valuable to have an overarching descriptive framework, a list of key components, and a clear idea of where and how students engage with these components and the entire process during their studies.

Porkess (2011) analysed the statistics curriculum for schools in England and Wales and concluded that there is no problem analysis and no data collection in the curriculum for pupils aged 11 years and above. Typically, students are told that a particular distribution can be applied to a data set and are then asked to demonstrate competence in performing a routine calculation. The inferential argument provides an underpinning intellectual theme. This problem is not confined to England and Wales: Velleman (1997) argues that statistics curricula too often teach techniques for data analysis without teaching why or how statisticians should use them. The future also looks bleak. In an analysis of proposed reforms to the statistics curriculum, the Royal Statistical Society argues that the curriculum in England and Wales 'will mainly test students on individual data presentation techniques. It continues to do little to assess students' statistical problem-solving abilities. Consequently we fear that the teaching of these skills will be neglected, leaving students without the adequate statistical skills they need for understanding and applying problem solving approaches in many other subjects'.

New Zealand provides an interesting contrast. New Zealand renamed the school mathematics curriculum *Mathematics and Statistics* in 2008. There is an emphasis throughout the curriculum on the usefulness of statistics—using statistics to solve problems that involve real (and interesting) data. A high proportion of students continue to study statistics beyond compulsory schooling; the New Zealand take-up rate for advanced mathematics is three times that of England. The New Zealand curriculum requires that pupils *inter alia* plan and conduct investigations using the statistical enquiry cycle, justify the variables and measures used, identify and communicate relationships between variables and differences within and between distributions, using multiple displays, make informal inferences about populations from sample data, justify findings, evaluate statistical reports in the media, use multiple displays and re-categorise data to find patterns, variations, relationships and trends in multivariate data sets.

In the UK and elsewhere, too much curriculum space is devoted to drawing conclusions about populations from small samples. This is easy to understand in an historical context where there was very little data on which to base decisions and where there was no computing power to handle large data sets. However, an exclusive focus on data from small samples has some obvious disadvantages. Estimates derived from small samples are unstable: students have to cope with the convoluted logic of hypothesis testing (see Gliner, Leech, & Morgan, 2002, for examples of errors in textbooks; Haller & Krauss, 2002, for instructor errors and Sotos, Vanhoof, Van den Noortgate & Onghena, 2007, for a review of student misconceptions); students need to understand the idea of the power of a test. Simulations can be used to show the instability of different estimates (and of p values in particular) as a function of effect size and sample size (see Cumming (2012) and Wild, Pfannkuch, Regan, & Horton (2011)) but are rarely used, nor are they proposed in new curricula for England and Wales.

Data from small samples lend themselves readily to questions about (say) the difference between means, but not to robust estimates of effect sizes. Decision-making is usually based on estimates of effect sizes, and information about the costs of different treatments, and likely benefits. For almost all practical purposes, it is essential to know the magnitude of a difference or an association; significance is a necessary condition but is useless on its own (see Gorard, 2014, for a critique of the misuse and corrosive effects of hypothesis testing in educational research).

Generalisation is also a key problem when reasoning about data from small samples. The plausibility of the generalisation is linked directly to the context. In a study where 10 samples of sodium of known purity are burned and they all burn with a yellow flame, it is reasonable to conclude that sodium will burn with a yellow flame anywhere in the world, now, and forever. If a sample of a recently discovered species of 10 individuals have yellow fur, we are likely to be cautious about a generalisation to all members of the species. From the viewpoint of policy, using data from small samples is particularly risky; decisions will be very heavily dependent on strong assumptions about generalisability. Consider the case of educational attainment shown in Figure 1. If one had data from just Chinese students, one would conclude that there is little or no association between eligibility for FSM and attainment; if one had a representative sample of black or white children, one would conclude that there is a strong association between FSM eligibility and attainment.

Significance testing and technical exercises based on small samples are prominent in statistics curricula for historical reasons, which are now barely relevant. Curriculum time should be devoted to more important topics; the conceptual and technical aspects of significance testing can be dealt with adequately by introducing students to randomisation tests (e.g. Lock *et al.* (2013)).

6 The Data Revolution and the Statistics Curriculum

Statistical literacy is usually conceived of as a broad set of dispositions and skills that are brought to bear when an individual is offered an argument where some of the evidence is based on data (e.g. Schield, 2015, 'Statistical literacy is critical thinking about statistics in arguments'). A number of technical skills are an essential component of statistical literacy such as the ability to read graphs (and to be aware of potential problems associated with scales that do not start at zero, distortion of axes, using areas to represent scalar quantities and the like, see, e.g. Tufte, 2001). Technical skills are the focus of most curricula, and problems at school, university and in the populace have been studied extensively (see http://www.statlit.org/ for a bibliography). Influential documents such as the Guidelines for Assessment and Instruction in Statistics Education reports produced by the American Statistical Association (Franklin (ed.), 2007 and Garfield (ed.), 2010) emphasise statistical literacy and advocate *inter alia* the use of real data, the importance of conceptual understanding and the use of technology. These are all challenges to much current statistics teaching (the reports do offer support for teachers). These reports were written in the early days of the data revolution; here, some necessary extensions are outlined, which are core components of statistical literacy.

Nicholson *et al.* (2006) and Ridgway *et al.* (2007a) argued that reasoning with data is pervasive in society but is largely ignored in the UK statistics curriculum. This problem has become more acute; the biggest problem with many current curricula is that they ignore the data revolution. The statistics curriculum faces two key challenges, ensuring that students engage with every phase of statistical problem solving and making students aware of current uses of statistical methods that affect their lives directly. Paradoxically, the data revolution can help students to experience the excitement of an earlier, less cluttered, statistical era. Many statistics courses

are focused on methods developed in the pre-computer era of the 1930s. Then, methods had to be developed that were tractable without computer power, and so made simplifying assumptions that are now no longer necessary (Cobb, 2007). Since then, there has been a great increase in methods for data analysis (Breiman, 2001). Big data in particular are associated with the creation of new classes of models such as those used for statistical learning (e.g. Hastie *et al.* (2009)). One can argue that a wider variety of techniques should be taught and (paradoxically) that a smaller number of techniques should form the mathematical core, in order to make the curriculum relevant to everyday reality by devoting more curriculum time to working on real problems and to addressing core statistical ideas.

The data revolution demands an understanding of core statistical ideas. Open data and big data can provide vivid illustrations of these core ideas. The following sections provide examples of core ideas that can be illustrated via big data and open data.

6.1 Politics of Measurement

Every choice of measures reflects value systems, purposes and technical issues. Examples of the influence of value systems are provided by contrasting the Organisation for Economic Cooperation and Development initiative *Beyond GDP* (underpinned by the Stiglitz Report (2009)), with publications from the International Monetary Fund (IMF). *Beyond GDP* advocates the use of a raft of measures of social progress and assessment of the state of the environment (such as changes in the stock of renewable and non-renewable resources during a government's stewardship) in holding governments to account. The recent crisis in the global financial market and the choice of measures used to reflect 'recovery' by the IMF and some governments (almost all financial measures) illustrate the way that measures reflect values and steer political decision-making in important ways.

The UN's MDG were adopted in 2000 as an attempt to commit governments worldwide to reduce poverty and promote well-being by measuring factors such as extreme poverty and hunger, disease incidence, primary education and gender equality, and thereby tracking governments' successes in improving performance on these measures. They provide an example of how well founded statistical systems might be used for social progress, by measuring key variables and holding governments accountable to public opinion worldwide. Because tracking MDG was entrusted to agencies that are vulnerable to political pressures within the UN, the redefinitions and methodological revisions of the MDG also provide an important example of political interference and the need for statisticians to be free from such interference (Pogge, 2010).

6.2 Measuring

An enquiry by the Public Administration Select Committee (2014) in the UK concluded that there is strong evidence that the police under-record crime, particularly sexual crimes such as rape. The UK Statistics Authority subsequently removed police-recorded crime data from its list of official national statistics. Data collection and measurement lie at the heart of statistical thinking. Core ideas are sampling, reliability and validity, each of which has many components (for example, concurrent validity, construct validity, face validity, predictive validity…). The key idea underpinning validity is to make the important measurable, not the measurable important. Good measures assess something we actually want to measure, with a degree of accuracy appropriate to the purpose the measure is designed for. Open data present some challenges, and the existence of detailed metadata makes the problems easier to see but does not make them go away. Return to the example of educational attainment in Figure 1, where attainment data are

presented for children entitled to FSM. In general (but not always), students who are eligible for FSM, on average, perform worse on measures of educational attainment than those who do not. So what is 'eligibility for FSM' a measure of? Can it be used as a surrogate for poverty? Or social class? Almost all measures in social science are problematic; 'health', 'poverty', 'inflation' and 'unemployment' all pose challenges. There is a need to engage students in the process of 'constructing and critiquing measures' (Ridgway, Swan, & Burkhardt, 2001).

The nature of big data means that extreme caution needs to be taken in its use. Hand (2013) points out that issues of data quality do not go away just because the data is present in huge volume. As well as the definition of measures (for example, the keywords used in sentiment analysis), there are interesting questions around sampling bias—these are obvious in the use of *twitter* posts to judge public happiness, issues of current concern and the like. Traffic sensors in the Netherlands provide a less obvious example. They generate 80 million records each day; 10 000 detection loops record the number of vehicles of different lengths passing, on a minute-by-minute basis. The number of big trucks is used as an indicator of economic activity; sensor data are used for road planning. However, coverage is patchy and does not represent the country as whole; the location of sensors introduces bias, because more sensors are placed at traffic intersections; malfunction of sensors adds another source of error. Data are extremely volatile. Boston's Street Bump smartphone app provides another traffic-related example. The app detects potholes in the road, and the data are used to plan road repairs. Such data are likely to over represent areas with a higher proportion of smartphone owners.

From the viewpoint of pedagogy, one can raise the issue of estimating the robustness of the evidence. Knowing about bias via unrepresentative sampling is one thing; estimating the likely size of the error is another. A key question is to ask if the data are good enough for the uses it is put to. Following Box & Draper (1987), one can ask how biassed data can be, before it becomes useless. One approach is to recommend that students look for some means of triangulation—for example, how does data from sensors relate to other sorts of data such as *post hoc* analyses of traffic jams?

The Internet makes it possible to use webscraping to provide estimates of variables such as the consumer price index, inflation or unemployment. These data could be used to calculate changes over very short time intervals. To do so is self evidently silly. However, deconstructing 'silly' is useful because it raises the issue of the relationship between the stability of some phenomenon and the accuracy of the measure used to describe it. It is common for journalists and politicians to offer rich interpretations of changes in key economic variables (or the popularity of political parties) over short periods of time that are well within the usual range of background variation in the measure used, such as monthly releases of data on unemployment or inflation.

The data revolution provides particularly poignant examples of Goodhart's law (1975). Goodhart's law (1975) refers to the idea that once a social or economic indicator is made a target for some aspect of policy, it soon ceases to measure what it used to measure. This happens because stakeholders find ways to raise or lower scores on the indicator via superficial, rather than deep, changes in system activities. One example is the introduction of waiting time as a measure of the performance of accident and emergency (AE) departments. Waiting time was defined to be the time difference between arriving in AE and being seen by a clinician. As a result of this choice of measure, at busy times, in some places, patients were kept waiting in ambulances outside the hospital (with a knock-on effect on ambulance services). Big data provide examples of systems actually designed to corrupt existing measures. Websites such as Swenzy (http://www.swenzy.com/) sell 'likes', 'downloads', *twitter* 'followers' and *Instagram* 'friends', which are actually bots (lines of code) written to simulate human activities on media websites. These are used to make companies, political parties and individuals appear more popular than they really are. Bilton (2014) reports that bots have been used extensively in political

campaigns in Syria, Turkey and Mexico. Bilton (2014) also reports that as a class demonstration at the Technion in Israel, students created bots in order to engineer a fake traffic jam on Google's satnav software, which led to traffic being diverted around the imaginary jam.

6.3 Data Exploration

The SMART Centre provides a number of interactive displays on topics such as educational attainment, health and riots, along with a facility for users to embed their own data in interactive multidimensional displays. Figure 1 provided an example. The Data Visualisation Centre within the UK Office for National Statistics (ONS) has created a variety of exciting data displays such as dynamic population pyramids, a dynamic map of commuter flow and series of choropleth maps to display small area census data. Figure 2 shows commuting patterns in the UK and allows users to scroll over the map to explore patterns in ways that are almost impossible simply using the underlying spreadsheets (for example, people commute from North Wales to England, and from South Wales to England, but rarely between North and South Wales).

Figure 3 shows the *constituency explorer*. Data synthesised from different sources are available on over 150 variables for every constituency (i.e. electoral districts for which members of parliament are elected) in the UK, in a series of linked displays. Individual constituencies can be identified and compared; metadata are available; data can be downloaded; the display runs on smartphones and laptops; there are links to associated print materials.

Data visualisation presents opportunities and challenges to statistics educators. The opportunities are clear—there are novel ways to present large volumes of complex data in ways

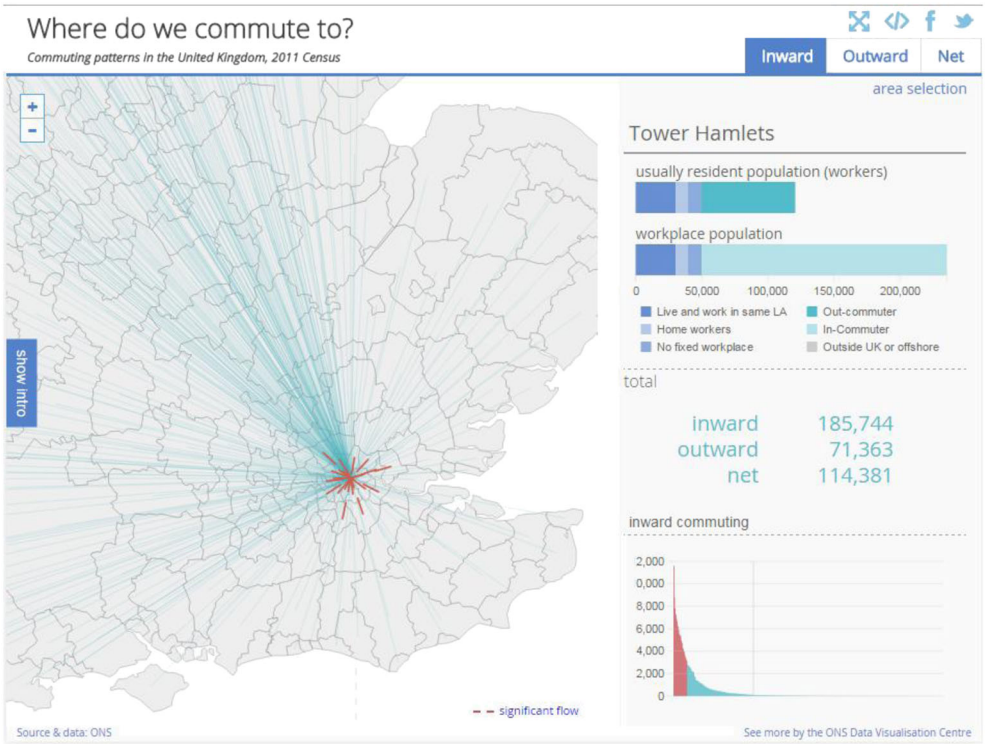

Figure 2. *Commuting patterns in the UK.*

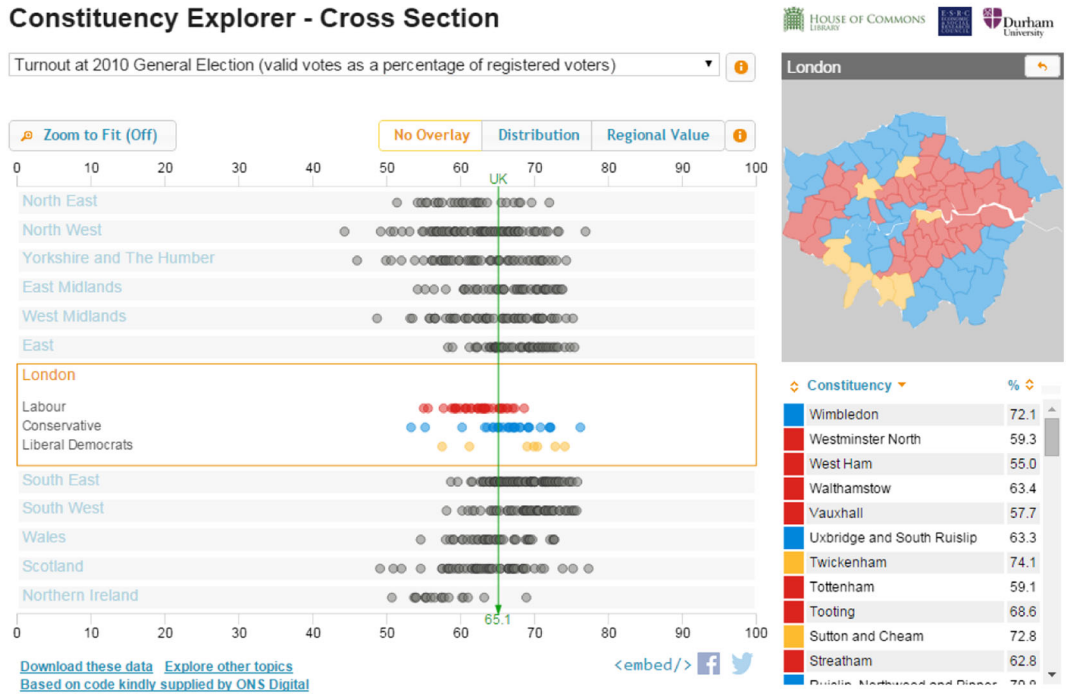

Figure 3. *The constituency explorer.*

that facilitate exploration by users—both expert and novice. For example, interactive maps of 'migration' and 'internal migration' offer users simple ways to explore complex patterns that are impossible to understand using just the underlying spreadsheets. User competence in working with data visualisations is largely undetermined (although there is some evidence that statistically naïve users can make sense of multivariate data presented in interactive displays—Ridgway, Nicholson, & McCusker, 2007b). Data visualisations hold the promise of direct access to rich, authentic and contemporary content to support exploration and decision-making. There are also challenges. Data visualisations often present multivariate data; interpretation of even simple line graphs can be problematic (e.g. Swan & Phillips, 1998). The variety of data visualisations continues to increase (see, for example, D3 library and Manyeyes), and so educators face the challenge of teaching generic skills in interpreting novel visualisations, rather than mastery over a small set of graphical conventions.

6.4 Association and Causality

The website http://www.tylervigen.com/ offers a new spurious correlation every day, by using open data sources and simple data-mining techniques. Examples include a correlation of $r = 0.99$ between US spending on science space and technology and suicides by hanging, strangulation and suffocation (1999–2009), and a correlation of $r = 0.96$ between US per capita consumption of mozzarella cheese and US civil engineering doctorates awarded (2000–2009). It illustrates the power of data mining in finding associations between variables and the foolishness of confusing correlation and causality. The interpretation of association is problematic in all branches of science. For example, drinking red wine has been associated with reduced incidence of heart disease. However, social class is a strong predictor of heart disease. It is

possible that a re-examination of the data might show that the effects of reading the Times or the New York Times have been as powerful as drinking red wine in prophylactic coronary care.

Big data also provoke a review of the 'correlation is not causation' mantra, in part, because the status of causality is quite different in social science and business. In social science, causality is important both theoretically and practically. Many people are concerned with understanding which factors are causal and which ones are not. From the viewpoint of practice, where policy changes are planned in (say) health, education or poverty relief, it is important to be confident that changes in policy are likely to result in changes in outcome. In business, a common strategy is to use information about association as the basis for action. Books (and other goods) are recommended on the basis of purchasing patterns. Information about which pairs of items have often been bought together in the past is used to advertise products to people who have purchased just one of the pair. Similarly, customer profiles are used to target advertising. There is little or no interest in causality; choice on one item (or being a member of a certain social group) does not cause future purchases, but using knowledge about association can lead to changes in behaviour that are useful to marketers (and perhaps to shoppers via information that 'other customers who bought <x> also bought <y>'). So causality is not a necessary condition for effective action, especially where the costs of action are modest and the consequences of error are minor.

Causality is a problematic concept to be explored in the curriculum. One can ask if there is an association there at all or if the observed data could reasonably be attributed to an underlying random process. One might offer a neat description of the data and pattern, for example, via mathematical models such as correlation coefficients, regression equations or odds ratios. In addition to association, plausible stories about observed patterns, invoking causal stories, can be created. These involve notions such as experimental intervention and control. One can critique causal stories, offer and where possible explore alternative 'third variable' accounts. The invention of critical tests of alternative explanations is an important part of science and should figure more strongly in the statistics curriculum (Pearl, 2009). We know rather little about how to scaffold such a curriculum. Pearl's work (e.g. Pearl, 2010) provides a starting point. There is a need to introduce probabilistic causality (if I smoke but do not get cancer, does this mean that there is no causal link?) built on simpler probabilistic ideas (Offered a choice of a $1 pay-off if a die comes up 1 or 2, or a $1 pay-off if it comes up 6, I choose the first bet. Wrong! It's a six. Should I change my strategy?).

6.5 Multiple Comparisons

Finding spurious correlations via data mining is a specific example of the multiple comparisons problem that besets much of conventional research in social science and medicine. The multiple-comparisons problem arises when a researcher conducts a large number of analyses to investigate the effect of some treatment. An example is the introduction of breakfast clubs into schools, where students are provided with a meal at the start of the day. Are breakfast clubs beneficial? Comparisons could be based on students of different ages and abilities and social background, girls and boys. Measures could include a large collection of academic measures, along with measures of social functioning and physical measures such as height, weight and incidence of different diseases. As the number of comparisons increases, so too does the probability of a spurious effect being detected. Calculating large numbers of correlations runs the same risk, as do exploratory investigations of data sets where there are large numbers of variables. Ioannidis (2005) wrote an article entitled *Why Most Published Research Findings Are False* to make this point. Again, open data and big data techniques provide examples that vividly illustrate well-known problems. While there are some technical corrections for multiple comparisons

problems where the number of comparisons being conducted is known, students should be presented with the basic challenge of linking evidence and theory in a meaningful way.

6.6 Fitting Data, Explaining and Predicting

Ginsberg *et al.* (2009) reported the success of GFT. By analysing 50 million search terms related to flu symptoms and the location of pharmacies, it was claimed that GFT was able to track the spread of influenza across the USA accurately, cheaply and far faster than the tracking methods used by the Centres for Disease Control and Prevention (CDC) that depended on surveillance reports from laboratories across the USA. GFT was a classic application of big data methodology—use exhaust data (web searches) that are plausibly related to the phenomenon of interest and that have a spatial component and look for predictors. GFT set out to identify the best predictors amongst 50 million search terms for 1 152 data points from CDC. In the case of flu, there are some obvious covariates such as time of year; GFT found that terms relevant to high school basketball (played in the winter months) were good predictors but took them out. GFT proved to be an effective predictor for the first 3 years of its use; then, it completely failed to predict the non-seasonal 2009 pandemic. In 2012, it produced an estimate that was double the CDC estimate and had been persistently overestimating the incidence of flu for a long period (Lazer, Kennedy, King, & Vespignani (2014)). Goel *et al.* (2010) showed that the predictions from GFT were little better than predictions from simple-lagged CDC data; Lazer *et al.* (2014) showed that lagged models of CDC data gave better predictions than GFT and that a combination of the two sets of data gave better predictions still.

The example is useful for statistics educators. It demonstrates the danger of overfitting—it is easy to fit a small number of data points with a large number of predictors. (Conventional practice would be to fit a subset of the data in order to build a model and then to test the model on a different subset). It demonstrates the danger of big data hubris. It illustrates the predictive power of past performance for new performance. It raises important questions about how models can be compared. Finding that big data has some predictive value is interesting; knowing how well these predictions compare to predictions derived in other ways (and the relative costs of different methods) is essential for theory and practice. A further problem is not knowing quite what the algorithm was actually doing; transparency is a useful property for a model to have.

The example of GFT is a reminder of the instability of models derived via conventional statistical methods. For example, the Framlingham risk equation was the equation of choice to predict the probability of a patient suffering a coronary vascular accident in the next 10 years, as a function of sex, age, cholesterol, smoking and other factors. In 2010, the National Institute for Clinical Excellence withdrew its recommendation that it be the equation of choice and proposed that it be considered as one of the possible equations to use (British Medical Association & the Royal Pharmaceutical Society (2014)). A possible cause is the rise in obesity in the population. Weight was not taken into account in the original sample used to determine the model parameters. Obesity predisposes one to coronary vascular accidents and a variety of other medical conditions. Unsurprisingly, changes in the shape of the population require new models. A related issue is the use of models derived primarily from data about one ethnic group (white) when dealing with people from other ethnic groups; again, warnings against this are a relatively recent phenomenon.

6.7 Combining Evidence

Access to large collections of research papers in a searchable form makes systematic literature reviews a powerful tool for shaping policy decisions in areas such as medicine

(the Cochrane collaboration) and in education, crime and justice, social welfare and international development (the Campbell collaboration). Meta-analysis provides good examples of the problems of data quality and important measurement issues (for example, are the measures of educational performance used in different studies sufficiently similar to justify aggregating the effect sizes found?). Design issues come to the fore—under what circumstances (if ever) is it sensible to combine data from randomised controlled trials with data from naturally occurring experiments? Students are exposed to statistical techniques estimating an overall effect size from studies that use different sample sizes; funnel plots raise awareness of the relationship between sample size and the stability of an estimate; funnel plots also point to the accidental suppression of results (because journals are unlikely to publish results from studies based on small samples where the results show no statistically significant effect) and to the deliberate suppression of results (where drug companies suppress the results from trials that do not demonstrate the effectiveness of their product).

Open data facilitates the synthesis of data from different sources. Data.gov has a facility to gather data from different government databases. Increasingly, there are efforts to synthesise big data and open data to improve predictive power such as the combination of GFT and CDC data by Goel *et al.* (2010).

Low response rates are a particular problem in survey research, in part because response rate is associated with a range of social indicators such as poverty. The Italian National Institute of Statistics (ISTAT) is developing an interesting approach to gathering census data. ISTAT is surveying a small group of deprived areas in detail, using repeated visits in order to get a high response rate. Mobile phone data is then used to map the relationships between demographics and phone activity. When these patterns have been established, they will use mobile phone data to complement survey data from areas of similar levels of deprivation.

6.8 Ethics

Access to data and data ownership are both contested. The open data movement has had very positive effects by increasing access to data sets collected by national and international agencies; however, large corporations (and government agencies such as the US National Security Agency) gather data on transactions of various sorts that may be commercially sensitive or related to issues of national security, and so data collectors resist allowing access to outside groups.

With open data, there are difficult issues around data release and anonymising data. Privacy issues are not easy to address. In the UK, there was a government plan to assemble medical records from local practitioners, which could be sold to private companies. These records were to be anonymised. Data would be provided about the postcode (zip), sex and age of patients, but nothing more. What could go wrong? In the case of the author, there are just two houses with the same postcode, each housing a heterosexual couple where the partners are about the same age; the ages of the two couples differ by about 10 years. One hardly needs a Rossetta Stone to work out exactly which person corresponds to which record.

With big data, a current problem is that governments and private companies are collecting information about citizens for their own advantage in ways many people might not know about, nor approve of. There is too little transparency about what is being collected, the reasons for the data collection or the uses to which they are actually being put. There is a range of important issues surrounding the legality and morality of data gathering that should be considered within the statistics community. Legal frameworks differ between countries. In the UK, for example, the *midata* initiative encourages companies to give customers

access to the data collected about them. Students could be asked to explain the different moral principles that underpin different laws.

7 Revising Statistics Curricula

How should statistics curricula be revised in the light of the data revolution? Nolan and Temple Lang (2010) make the case that at undergraduate level, students should be taught appropriate skills in statistical computing and point to the broad array of computational topics relevant to statistics (p102). One might argue that if statisticians do not acquire appropriate computing skills and further develop the field of statistical computing, they will be increasingly marginalised in the commercial and academic worlds; computer scientists will acquire or invent the statistics they need to engage with big data and open data.

At all levels of statistics education, the simple message is that the curriculum should be focussed on the development of statistical thinking and that the prime sources of data and phenomena to be modelled should be drawn from the data revolution. Students need to learn strategies for tackling open-ended, fuzzy questions in the context of rich data resources. Some recommendations follow.

Devote more space to open data: it is self evident that a lot of rich high-quality data are likely to be more useful than a small sample of high quality data. Open data let students see the role of careful data collection and analysis in understanding phenomena and in policy making. Open data are particularly useful in providing a rich picture of *what is*. Differences between subgroups are immediately obvious, and the size of the differences can be judged directly. Patterns in multivariate data allow one to judge the likely effect of policy initiatives on different subgroups such as people in different regions, different levels of educational attainment or different demographic characteristics.

Introduce multivariate data early: paradoxically, ideas that seem difficult when introduced late in the curriculum can be assimilated easily when introduced early. For example, the idea of non-linearity is only difficult because many statistical techniques assume linearity. Assuming linearity would be laughable in school science. Similarly, the idea of interaction is difficult because students are introduced to one and two variable problems and then face a seemingly difficult idea when working with three variables. By starting with multivariate data, interaction is an obvious feature of the data landscape. An example is provided in Figure 4, which shows the incidence of new episodes of STD from Genito—urinary medicine clinics. The data show interactions between disease type, age, sex and time.

One might expect such a complex interaction to be unintelligible, but the patterns are quite clear to see when students explore the data in an interactive display (e.g. Ridgway & Nicholson (2010) have used this display with statistically naïve 14 year-old students). The incidence of chlamydia shows a dramatic increase over a short period of time; other STD do not; women have contracted their first STD earlier than men; increases in incidence are different for different age groups. The data have clear implications for policy, both in terms of urgency and target group.

Teach about, and with, interactive graphics: there is a large and growing collection of interactive data visualisations designed to support the teaching of statistical concepts. Increasingly, these are available at no cost, for example, from the Open University, do not require proprietary software and run on mobile devices. These displays can support concept development and should be used more widely. Students should learn to read and critique novel data visualisations.

Work with multiple data sources: the *constituency explorer* provides an example where data from different sources are assembled. Meta-analysis can be used to highlight a range of questions around research quality, metadata issues and techniques for synthesising results. More

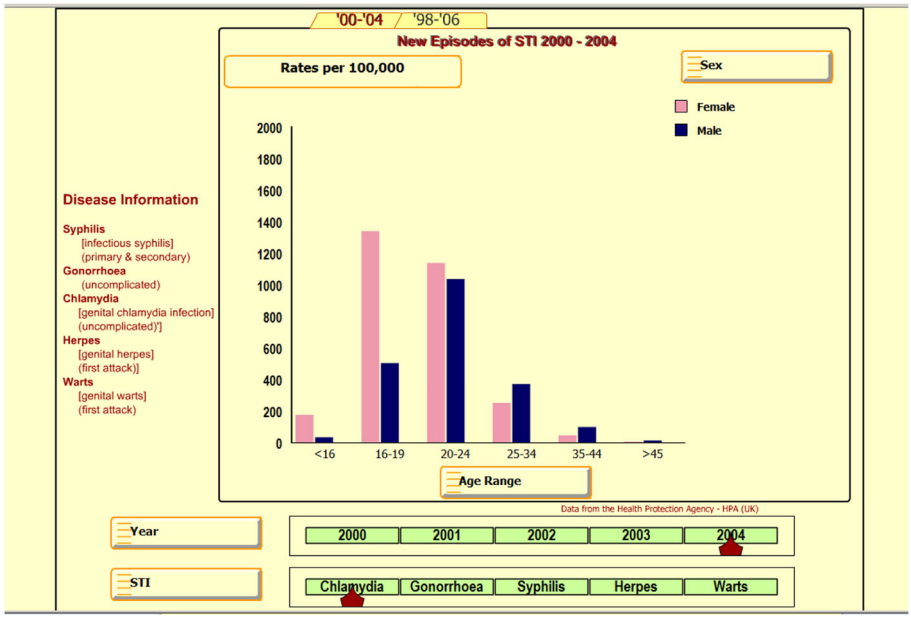

Figure 4. *New episodes of sexually transmitted disease.*

complex challenges can be to synthesise open data and big data and to use (cheap) big data as a surrogate for (expensive) open data.

Teach with, and about, familiar technologies: smartphones, fitness trainers and tablets offer opportunities to provoke statistical thinking. Fingerprint, voice and face recognition are commonplace; students can be challenged to explain Bayesian pattern recognition. Data from devices such as *fitbit* and phones' locations provide high volume, noisy data streams. Students can be asked to explain their own activity patterns over a time period or relate their activity to people with similar and different characteristics. Targeted advertising via the Internet can be used as a focus to ask technical questions about how to do it and ethical questions about the morality of overt and hidden surveillance. In-store advertisements on digital video screens based on face recognition, in-store movement-tracking cameras, mobile phone tracking, information collected via sign ups to free in-store wi-fi and location-specific messages sent to mobile phones based on past purchase history (e.g. discount vouchers) all raise both technical and ethical questions.

Engage with modelling: the data revolution provokes debate on the nature and purposes of modelling and the whole process of scientific enquiry. Modelling itself should be a focal curriculum topic. Students need to experience creating and critiquing measures. Students should model the same phenomena in different ways and different phenomena using the same formalism. Critique of different models should be encouraged, as part of a process of developing an aesthetic of modelling. Students should learn from failures by examining models that failed and studying the history of models that were initially successful then lost their predictive power. The politics of data and of modelling should be explored.

Use Internet resources to invigorate teaching: a number of sources generate provocations to statistical thinking on a regular basis, such as Vigen's website that provides a new (spurious) correlation each day; flowing data (https://flowingdata.com/) distributes novel data visualisations, often on statistically relevant topics; full fact (https://fullfact.org/) checks the assertions made

in media on a daily basis and the New York Times circulates summaries of technology-related articles (with hyperlinks), often relevant to the data revolution. These can be used as occasional foci of teaching; students can be encouraged to engage with these sources as extra-curricular activities.

Place more emphasis on decision-making and risk: examine the uses of evidence in decision-making. Higgins *et al.* (2014) link effect size, costs and the robustness of evidence in a review of meta-analyses in education. Evidence often informs decision-making but is sometimes invented by politicians as a warrant to justify decisions that have already been made, or by media to support ideological positions. Full fact monitors statements in the public domain (e.g. from politicians and journalists) and reports on inaccuracies; their website provides interesting examples of data misuse.

Illuminate current curriculum content with examples from the data revolution: the earlier section on statistical literacy highlighted the need to increase the attention paid to data provenance and data quality. The data revolution provides challenges regarding the definition and measurement of concepts and of sampling and bias (for example, using any technology as a source of evidence will result in data samples that are not representative of the population as a whole). Commercial uses of big data show that sometimes, correlation is enough for effective decisions and actions (e.g. recommending products to customers); other uses highlight the problems of atheoretical decisions and actions (e.g. GFT overfitting data). Data mining highlights the problems of making multiple comparisons.

Decrease the time allocated to hypothesis testing: far too much curriculum time is devoted to hypothesis testing; statistical significance is almost guaranteed with large samples. At school level, hypothesis testing should be replaced with randomisation tests; at the undergraduate level, emphasis should be placed on simulations that demonstrate the instability of p values with small samples, and hypothesis testing should be contextualised via discussions of confidence intervals, power and effect size.

8 Barriers to Change

There are barriers to overcome if curriculum reform is to be effective. First is inertia—it is easier to continue with current practices than to engage in the practice of constant revision. Teachers are invested in what they know—the linear model has a dominant place in the curriculum for good reason, and fostering student understanding currently takes up all the available time. Teachers need to acquire both new content knowledge (e.g. statistical computing, big data techniques and pattern recognition) and more pedagogical content knowledge (e.g. how to impart modelling skills). Acquiring new skills is effortful; a bigger problem is to teach about the revolution during the revolution.

There are structural barriers to change, in the form of existing curriculum structures, and assessment systems. It is easier to assess technical mastery than statistical thinking. There is also a sea of ignorance around the difficulty of concepts associated with the data revolution and logical connections between them, which will make curriculum planning difficult.

There are considerable costs associated with computer use, especially in early grades, and even higher costs associated with professional development.

9 Conclusions

Statistics educators need to respond positively to the opportunities provided by the data revolution. The alternative is to see the increasing irrelevance of a static statistics curriculum that

offers little help in understanding the data (and the ways that data are used) that affects the life of everyone. Changes need to be quite radical; new ways of approaching evidence should be adopted. However, this radicalism is entirely consistent with the ambitions of the founders of statistics and with current conceptions of statistical thinking and statistical literacy.

Acknowledgement

This work was supported by the Economic and Social Research Council (grant number ES/K004328/1).

References

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. Available at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory Accessed 12 May 2015.

Bardi, U. (2008). Cassandra's curse: how 'The Limits to Growth' was demonized. Available at http://www.theoildrum.com/node/3551. Accessed 12 May 2015.

Batanero, C., Burrill, G. & Reading, C. (eds). (2011). *Teaching Statistics in School Mathematics - Challenges for Teaching and Teacher Education: A Joint ICME/IASE Study.* Heidelberg: Springer.

Bilton, N. (2014). Friends, and influence, for sale online. Available at http://bits.blogs.nytimes.com/2014/04/20/friends-and-influence-for-sale-online/?emc=edit_ct_20140424&nl=technology&nlid=66226932. Accessed 12 May 2015.

Box, G. & Draper, N. (1987). *Empirical Model-Building and Response Surfaces.* New York: Wiley.

Bradshaw, P. (2010). How to be a data journalist. Available at http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide. Accessed 12 May 2015.

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.*, **16**(3), 199–231.

British Medical Association & the Royal Pharmaceutical Society. (2014). *British National Formulary* London: BMJ Group.

Brooke, H. (2010). *The Revolution will be Digitised: Dispatches from the Information War.* London: Heinemann.

Chatfield, C. (1991). Avoiding statistical pitfalls. *Stat. Sci.*, **6**(3), 240–268.

Cobb, G.W. (2007). The introductory statistics course: a ptolemaic curriculum? *Tech. Innovat. Stat. Educ.*, **1**(1), 1–15.

Condorcet, J. (1994). *Foundations of Social Choice and Political Theory.* Aldershot and Brookfield, VT: Elgar (original work published in 1792).

Cumming, G. (2012). *Understanding the New Statistics.* New York: Routledge.

Franklin, C. (ed). (2007). Guidelines for assessment and instruction in statistics education (GAISE) report:a PreK-12 curriculum framework. Available at http://www.amstat.org/education/gaise/. Accessed 12 May 2015.

Gal, I. (2002). Adult statistical literacy: meanings, components, responsibilities. *Int. Stat. Rev.*, **70**(1), 1–25.

Garfield, J. (ed). (2010). Guidelines for assessment and instruction in statistics education (GAISE) college report. Available at http://www.amstat.org/education/gaise/. Accessed 12 May 2015.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. & Brilliant, L. (2009). Detecting influenza epidemics using seach engine query data. *Nature*, **457**, 1012–1014.

Gliner, J.A., Leech, N.L. & Morgan, G.A. (2002). Problems with null hypothesis significance testing (NHST): what do the textbooks say? *Journal of Experimental Education*, **71**(1), 83–92.

Goel, S., Hofman, J., Lahaie, S., Pennock, D. & Watts, D. (2010). Predicting consumer behavior with web search. *PNAS*, **107**(41), 17488–17490.

Goodhart, C.A.E. (1975). Monetary relationships: a view from Threadneedle Street. Papers in Monetary Economics (Reserve Bank of Australia) I. cited in http://en.wikipedia.org/wiki/Goodhart%27s_law. Accessed 12 May 2015.

Gorard, S. (2014). The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward? *Psychology of Education Review*, **38**(1), 3–10.

Gray, J., Chambers, L. & Bounegru, L. (2012). *The Data Journalism Handbook.* Sebastapol, CA: O'Reilly Media.

Grolemund, G. & Wickham, H. (2014). A cognitive interpretation of data analysis. *Int. Stat. Rev.*, **82**(2), 184–204.

Hahn, G. & Doganaksoy, N. (2011). *A Career in Statistics.* New Jersey: Wiley.

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Meth. Psycho. Res. Online*, **7**(1), 1–20.

Hand, D.J. (2009). Modern statistics: the myth and the magic. *J. R. Statist. Soc. A*, **172**, 287–306.

Hand, D.J. (2013). *Big data hope or hype*. Available at https://www.knaw.nl/shared/resources/actueel/bestanden/20130910_bigdatasciencepresentatie_davidhand.pdf. Accessed 12 May 2015.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Higgins, S., Katsipataki, M., Coleman, R., Henderson, P., Major, L.E. & Coe, R. (2014). *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit.* London: Education Endowment Foundation.

Huber, P. (2011). *Data Analysis: What Can be Learned from the Past 50 Years.* Hoboken, NJ: Wiley.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, **2**(8), e124.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions.* Chicago: Chicago University Press.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, **343**, 1203–1205.

Lock, R., Lock, P., Lock, K., Lock, E. & Lock, D. (2013). *Statistics: Unlocking the Power of Data.* London: Wiley.

Madigan, D. (2014). *Statistics and science: a report of the London workshop on the future of the statistical sciences*. Available at http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf. Accessed 12 May 2015.

Meadows, D., Meadows, D., Randers, J. & Behrens, W. (1972). *The Limits to Growth*. Available at http://www.donellameadows.org/wp-content/userfiles/Limits-to-Growth-digital-scan-version.pdf. Accessed 12 May 2015.

Neurath, O. (2010). *From Hieroglyphics to Isotype: A Visual Autobiography.* London: Hyphen Press.

Nicholson, J.R., Ridgway, J. & McCusker, S. (2006). Reasoning with data—time for a rethink? *Teach. Stat.*, **28**(1), 2–9.

Nolan, D. & Temple Lang, D. (2010). Computing in the statistics curricula. *Amer. Statist.*, **64**(2), 97–107.

Pearl, J. (2009). Causal inference in statistics: an overview. *Stat. Surv.*, **3**, 96–146.

Pearl, J. (2010). An introduction to causal inference. *Int. J. Biostat.*, **6**(2). ISSN (Online) 1557–4679, DOI: 10.2202/1557-4679.1203.

Pogge, T. (2010). *Politics as Usual: What Lies Behind the Pro-Poor Rhetoric.* Cambridge: Polity Press.

Porkess, R. (2011). The future of statistics in our schools and colleges. Available from: https://www.rss.org.uk/uploadedfiles/userfiles/files/The%20Future%20of%20Statistics%20in%20our%20Schools%20and%20Colleges.pdf. Accessed 12 May 2015.

Public Administration Select Committee. (2014). Caught red-handed: why we can't count on police recorded crime statistics. Available at http://www.parliament.uk/business/committees/committees-a-z/commons-select/public-administration-select-committee/news/crime-stats-substantive/. Accessed 12 May 2015.

Pullinger, J. (2014). Statistics making an impact. *J. R. Statist. Soc. A*, **176**(4), 819–839.

Ridgway, J., McCusker, S. & Nicholson, J. (2003). Reasoning with evidence—development of a scale. In *International Association for Educational Assessment Conference*, Manchester, UK, pp. 10. Available at https://www.dur.ac.uk/resources/smart.centre/Publications/ReasoningwithEvidence-Developmentofascale.pdf. Accessed 12 May 2015.

Ridgway, J. & Nicholson, J. (2010). Pupils reasoning with information and misinformation. Paper presented at ICOTS8. Available at https://www.dur.ac.uk/resources/smart.centre/Publications/ICOTS8_9A3_RIDGWAY.pdf. Accessed 12 May 2015.

Ridgway, J., Nicholson, J. & McCusker, S. (2007a). Reasoning with multivariate evidence. *IEJME*, **2**(3), 245–269.

Ridgway, J., Nicholson, J. & McCusker, S. (2007b). Teaching statistics—despite its applications. *Teach. Stat.*, **29**(2), 44–48.

Ridgway, J., Nicholson, J. & McCusker, S. (2011). Reconceptualising 'statistics' and 'education'. In *Statistics Education in School Mathematics: Challenges for Teaching and Teacher Education*, Ed. C. Batanero. Heidelberg: Springer.

Ridgway, J., Nicholson, J. & McCusker, S. (2013). 'Open data' and the semantic web require a rethink on statistics teaching. *Tech. Inno. Stat. Educ.*, **7**(2), 12.

Ridgway, J., Swan, M. & Burkhardt, H. (2001). Assessing mathematical thinking Via FLAG. In *Teaching and Learning Mathematics at University Level—An ICMI Study.* Eds. D. Holton & M. Niss, pp. 423–430. Dordrecht: Kluwer Academic Publishers.

Royal Statistical Society. (2014). Royal Statistical Society response to the DfE policy statement on 16 to 18 core mathematics qualifications in England. Available at https://www.rss.org.uk/site/cms/newsarticle.asp?chapter=_15&nid=_137. Accessed 12 May 2015.

Schield, M. (2015). Available at: http://web.augsburg.edu/~schield/.

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction.* London: Penguin.

Sotos, A.E.C., Vanhoof, S., Van den Noortgate, W. & Onghena, P. (2007). Students' misconceptions of statistical inference: a review of the empirical evidence from research on statistics education. *Educ. Res. Rev.*, **2**, 98–113.

Stiglitz, J., Sen, A. & Fitoussi, J.P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress. OECD. Available at http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf. Accessed 12 May 2015.

Swan, M. & Phillips, R. (1998). Graph interpretation skills among lower-achieving school leavers. *Res. Educ.*, **60**, 10–20.
Tufte, E.R. (2001). *The Visual Display of Quantitative Information*, 2nd edn. Cheshire, CT: Graphics Press.
Tukey, J.W. (1962). The future of data analysis. *Ann. Math. Stat.*, **33**(1), 1–67.
US Congress. (2004). Budget of United States Government Fiscal Year 2004. *House Document*, **3**(1), pp. 356. United States Congressional Serial Set 14820.
Velleman, P.F. (1997). The philosophical past and the digital future of data analysis. In *The Practice of Data Analysis: Essays in Honor of John W. Tukey.* Eds. D.R. Brillinger, L.T. Fernholz & S. Morgenthaler. Princeton: Princeton University Press; 317–337.
Wikipedia. (2014). Available at: http://en.wikipedia.org/wiki/Grounded_theory.
Wild, C., Pfannkuch, M., Regan, M. & Horton, N. (2011). Towards more accessible conceptions of statistical inference. *J. R. Statist. Soc. A*, **174**(2), 1–23.
Wild, C. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *Int. Stat. Rev.*, **67**(3), 223–248.
World of Statistics. (2014). *What is statistics?* Available at http://www.worldofstatistics.org/2012/12/03/what-is-statistics/. Accessed 12 May 2015.
Zetie, K.P. & James, J.E.M. (2002). Extreme value theory. *Phys. Educ.*, **37**, 381–383.

[*Received June 2014, accepted May 2015*]